ISPRS Int. J. Geo-Inf. **2015**, *4*, 471-492; doi:10.3390/ijgi4020471

OPEN ACCESS

ISPRS International Journal of
Geo-Information
ISSN 2220-9964
www.mdpi.com/journal/ijgi

Article

A Structural-Lexical Measure of Semantic Similarity for Geo-Knowledge Graphs

Andrea Ballatore [1],*, Michela Bertolotto [2] and David C. Wilson [3]

[1] Center for Spatial Studies, University of California, Santa Barbara, CA 93106, USA

[2] School of Computer Science and Informatics, University College Dublin, Dublin 4, Ireland;
E-Mail: michela.bertolotto@ucd.ie

[3] Department of Software and Information Systems, University of North Carolina, Charlotte,
NC 28223, USA; E-Mail: davils@uncc.edu

* Author to whom correspondence should be addressed; E-Mail: aballatore@spatial.ucsb.edu;
Tel.: +1-805-893-5267

Academic Editor: Wolfgang Kainz

Received: 24 December 2014 / Accepted: 23 March 2015 / Published: 1 April 2015

Abstract: Graphs have become ubiquitous structures to encode geographic knowledge online. The Semantic Web's linked open data, folksonomies, wiki websites and open gazetteers can be seen as geo-knowledge graphs, that is labeled graphs whose vertices represent geographic concepts and whose edges encode the relations between concepts. To compute the semantic similarity of concepts in such structures, this article defines the network-lexical similarity measure (NLS). This measure estimates similarity by combining two complementary sources of information: the network similarity of vertices and the semantic similarity of the lexical definitions. NLS is evaluated on the OpenStreetMap Semantic Network, a crowdsourced geo-knowledge graph that describes geographic concepts. The hybrid approach outperforms both network and lexical measures, obtaining very strong correlation with the similarity judgments of human subjects.

Keywords: geographic concepts; semantic similarity; geo-knowledge graphs; network-lexical similarity measure (NLS); lexical similarity; network similarity

1. Introduction

Computing the similarity of concepts in a knowledge-representation structure is a cornerstone for a wide variety of advanced tasks in geographic information science (GIScience), geographic information retrieval, geoparsing, natural language processing and artificial intelligence. In this article, we combine network similarity and lexical similarity measures into a hybrid measure in order to compute the similarity of geographic concepts in graph-based structures that represent geographic knowledge, showing empirically that both aspects of similarity increase the performance. Because of their simplicity and their adherence with human semantic intuition, graphs have been the most popular knowledge-representation structure over the past 30 years [1]. A wide variety of geographic knowledge bases rely on some form of graph-based representation, ranging from gazetteers, geo-databases, location-based social media and wikis, to the linked open data cloud that emerged from Semantic Web research [2] (http://lod-cloud.net).

At the core, these structures can be seen as geo-knowledge graphs (GKGs). In this article, we define a GKG as a representational artifact that contains geographic concepts, their mutual relations, and their lexical descriptions. GKGs do not necessarily attach formal constraints to their concepts and relations. Please note that GKG is a general term that does not refer to Google's Knowledge Graph (http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html). Formally, a GKG is a directed graph whose vertices correspond to concepts and whose edges are relations. Lexical descriptions of concepts are associated with vertices. Hence, such knowledge-representation structures are ubiquitous: even websites can be seen as GKGs, in which each page is a concept, and hyperlinks represent a generic, unspecified relation. More complex logic formalisms, such as conceptual graphs and ontologies, still contain GKGs.

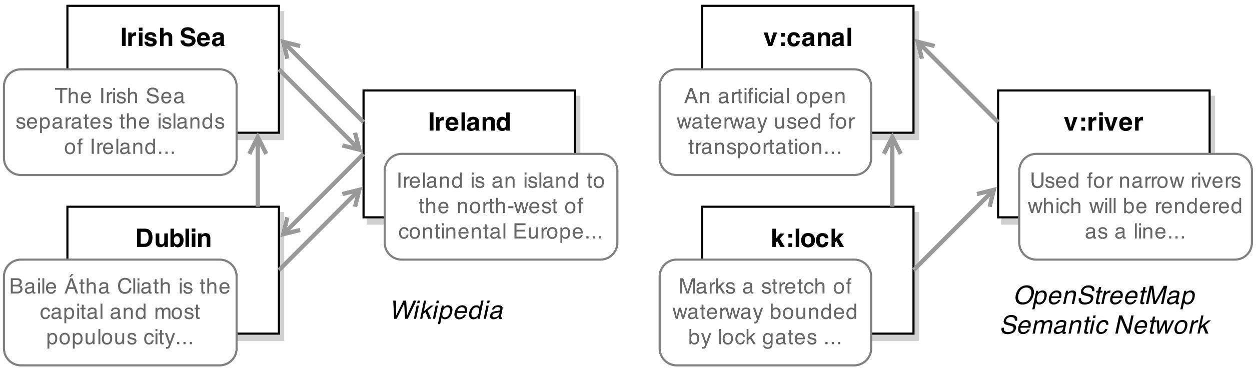
Irish Sea
The Irish Sea separates the islands of Ireland...
Ireland
Ireland is an island to the north-west of continental Europe...
Dublin
Baile Átha Cliath is the capital and most populous city...
Wikipedia
v:canal
An artificial open waterway used for transportation...
v:river
Used for narrow rivers which will be rendered as a line...
k:lock
Marks a stretch of waterway bounded by lock gates ...
OpenStreetMap Semantic Network

Figure 1. Fragments of geo-knowledge graphs (GKGs) extracted from Wikipedia and the OpenStreetMap (OSM).

Traditionally, knowledge-representation artifacts were built by experts for specific scientific or engineering purposes, such as the lexical database WordNet [3] and the artificial intelligence project Cyc (http://www.cyc.com). With the emergence and sophistication of volunteered geographic information (VGI) [4], GKGs are often characterized by highly variable coverage and quality [5]. GeoNames (http://www.geonames.org) can be seen as a GKG, in which the gazetteer entries are concepts connected through hierarchical and other relations. GKGs are found in the hyperlinked graph of Wikipedia articles, as well as in cartographic projects, such as OpenStreetMap, which focuses on spatial vector data (see Figure 1).

In the context of such knowledge-representation graphs, given two concepts in a GKG (or in different GKGs), a semantic similarity measure aims at quantifying their similarity as a real number, typically normalized in the interval $\in [0, 1]$. Typically, the computation of semantic similarity is not an end in itself, but is an intermediate task necessary to enable other tasks. For example, in a given GKG, the concepts "river" and "canal" might have similarity 0.75, whilst "river" and "restaurant" score only 0.05. If a measure mimics human judgment to a sufficient degree, these similarity values can be used for query relaxation in geographic information retrieval, including canals in the results for a query aimed at rivers, as well as for conceptual alignment, detecting high similarity between different representations of the concept river in two geo-databases [6,7].

In our previous work, we investigated the application of network-based and lexical similarity measures to compute the semantic similarity of geographic concepts, in the context of a crowdsourced semantic network [8–10]. The original contribution of this article builds upon and extends this body of work in several respects. First, we devise a hybrid semantic similarity measure, the network-lexical similarity measure (NLS), which combines two pillars of the similarity of concepts in GKGs. We define the first pillar as the concept's topological location, *i.e.*, its structural relations with other concepts. The second pillar is based on the semantic similarity of concepts' lexical definitions, expressed in natural language. NLS combines network and lexical similarity measures, and both aspects contribute to increasing the cognitive plausibility of the measure, *i.e.*, the ability of the measure to mimic human judgments. To the best of our knowledge, NLS is the first approach to semantic similarity that combines these two aspects.

Second, the cognitive plausibility of NLS is thoroughly evaluated on a real-world GKG, the OpenStreetMap (OSM) Semantic Network [5], which contains about 5000 concepts extracted from crowdsourcing project OpenStreetMap. This GKG allows a detailed assessment of NLS in the context of a geographic knowledge-representation artifact, enabling a critical discussion on the limitations of network and lexical similarity measures. As ground truth, this evaluation utilizes the geo-relatedness and similarity dataset (GeReSiD) [10], providing a more reliable and extensive evaluation, and allows a detailed comparison of the measures. These results are compared with those obtained with the dataset used to evaluate the matching-distance similarity measure (MDSM) [11]. The empirical results of this study further the assessment of the cognitive plausibility of network and lexical measures and confirm the high cognitive plausibility of NLS, which consistently outperforms both network and lexical measures.

The remainder of this article is organized as follows. Section 2 surveys relevant literature on semantic similarity. Section 3 outlines NLS, the proposed hybrid measure of semantic similarity. Subsequently, a detailed empirical evaluation of the measure is presented and discussed (Section 4). We conclude with a summary and discussion of directions for future research (Section 5).

2. Background

Semantic similarity is a specific type of semantic relatedness, based on subsumption relations (*is a*) [10]. For example, "fuel" is semantically related to "car", while "bus" is semantically related and similar to "car". Given the fundamental nature of semantic similarity, it is difficult to provide a definition without circularity, and several terms have been used to discuss it. "Semantic distance" is used to refer to the distance between two concepts represented in a geometric semantic model [12]. Depending on what attributes and relations are considered, semantic similarity can be computed as inversely proportional to semantic distance. Furthermore, the term "semantic association" is used to define semantic relatedness, in particular in human memory retrieval processes. "Taxonomical similarity", on the other hand, is equivalent to semantic similarity [13]. In a GKG, concepts are connected through relations that express their general semantic relatedness.

In the context of GIScience, measures of semantic similarity and relatedness are widely applied in geographic information retrieval, data mining and geo-semantics [6,14]. Specific measures of semantic similarity tailored to geographic concepts have emerged [15]. Rodríguez and Egenhofer [11] have extended Tversky's ratio model, taking context explicitly into account, by selecting a subset of features based on user needs. Janowicz *et al.* [14] have developed a similarity measure for geographic concepts based on description logic (DL), a family of Semantic Web languages. Such measures can only be applied to concepts expressed in specific formalisms, such as DL. Hence, in the context of GKGs, these measures are not directly applicable, and different approaches are needed.

2.1. Network Similarity Measures

This section describes existing techniques to compute the similarity of vertices in graphs, the first component of NLS. These approaches to similarity are based on some form of structural distance between nodes, such as edge counting, sometimes adding additional parameters to weight the paths [16]. Such network-based techniques have been applied to well-defined, expert-generated semantic networks in which the edges are expressed in some formal semantics, such as WordNet. However, the GKGs we are focusing on do not present such a semantically-rich structure, but encode knowledge in the form of simple graphs of inter-linked objects. Given the popularity of networks in many fields, several algorithms have emerged to identify similar objects exclusively on their link patterns in graphs that do not explicitly formalize relations.

Small [17] devised the seminal "co-citation" algorithm. Given a graph representing scientific articles and their mutual references, this measure models the similarity between two given papers by the frequency in which they are cited together. Extending co-citation to a recursive form, Jeh and Widom [18] created SimRank, an approach to calculating vertex similarity in directed graphs. The underlying circular intuition is that two objects can be considered similar if they are referenced by similar objects. The P-Rank algorithm [19] further extends co-citation by taking into account outgoing links. Previous network similarity algorithms, such as the original co-citation [17], Coupling [20] and Amsler [21], are specific cases of P-Rank. In our previous work, we showed that, when applied to geographic concepts, SimRank and P-Rank tend to reach higher plausibility than the other network measures [8].

2.2. Lexical Similarity Measures

The general objective of lexical similarity measures is the quantification of the similarity of two lexical units, typically as a real number. A lexical unit can be either an individual word, a compound word or a segment of text [22]. Approaches to compute the semantic similarity of individual words (as opposed to larger semantic entities) can be classified into two main families: knowledge-based and corpus-based. Knowledge-based techniques utilize manually-generated artifacts as a source of conceptual knowledge. Under a structuralist assumption, most of these techniques observe the relationships that link the terms, assuming, for example, that the ontological distance is inversely proportional to the semantic similarity [23]. WordNet [3] has been used to compute lexical similarity with a variety of methods, as shown in Table 1 [24–30]. These measures obtain varying plausibility depending on the context and can be combined into ensembles to obtain higher plausibility [31]. Corpus-based techniques, on the other hand, do not need explicit relationships between terms and compute the semantic similarity of two terms based on their co-occurrence in a large corpus of text documents [32,33].

Table 1. WordNet-based similarity measures. *lcs:* least common subsumer.

Name	Description	Name	Description
path [23]	Edge count	wup [24]	Edge count between lcs and terms
lch [25]	Edge count scaled by depth	hso [26]	Paths in lexical chains
res [16]	Information content of lcs	lesk [27]	Extended gloss overlap
jcn [28]	Information content of lcs and terms	vector [29]	Second order co-occurrence vectors
lin [30]	Ratio of information content of lcs and terms	vectorp [29]	Pairwise second order co-occurrence vectors

Semantic similarity can be computed between segments of texts, in a linguistic problem called "paraphrase detection". For example, the sentence "Any trip to Italy should include a visit to Tuscany to sample their wines" bears high semantic similarity with and is a paraphrase of "Be sure to include a Tuscan wine-tasting experience when visiting Italy". To tackle this issue, Corley and Mihalcea [34] developed a knowledge-based bag-of-words technique to paraphrase detection, which relies on some of the WordNet measures. In our previous work, we developed a similarity measure geared towards lexical definitions [9]. In terms of precision, the knowledge-based measures generally outperform the corpus-based ones [35]. Although numerous semantic similarity measures exist, to the best of our knowledge, no hybrid measure has been proposed that combines network and lexical similarity for GKGs. The next section outlines NLS, our approach to filling this knowledge gap.

3. The Network-Lexical Similarity Measure

The general problem that network-lexical similarity measure (NLS) aims at solving is the quantification of semantic similarity in a GKG. Formally, a GKG is a labeled graph $G(V, E, L)$, with a set of vertices V (concepts), a set of directed edges E (relations) and a set of labels L (lexical definitions). A labeling function $V \rightarrow L$ associates the vertices with labels. A label $l \in L$ contains a segment of text and can be empty. A directed edge $e \in E$ associates two nodes $e = \{u, v\}$, where $u, v \in V$. Given two concepts a and $b \in V$, the objective of a semantic similarity measure is to compute a similarity score $s(a, b) \in \Re$. To ease their interpretation, the scores are normalized in the interval $\in [0, 1]$. It is important to note that similarity scores are not meaningful in isolation, but convey useful information when compared with other pairs of concepts.

In order to compute the semantic similarity in GKGs, NLS rests on two pillars: network similarity s_{net} and lexical similarity s_{lex}. The network similarity of two concepts is extracted from their topological location in the graph, observing the link structure of G. On the other hand, the lexical similarity focuses on the labels in L that contain segments of text describing the concepts. Natural language processing techniques can thus be used to measure the semantic similarity of segments of text. These two perspectives on concept similarity are not mutually exclusive, and NLS considers them as complementary. NLS should be seen as a general framework to compute semantic similarity, combining complementary aspects of similarity in GKGs.

3.1. Network Similarity (s_{net})

The network similarity function $s_{net}(a, b)$ aims at quantifying the structural similarity of vertices in graph G. Because GKGs do not encode the formal semantics of the relations between concepts, suitable measures have to consider edges as indicators of general relatedness. If either a or b is not connected to other nodes, $s_{net}(a, b)$ is undetermined, and NLS relies only on $s_{lex}(a, b)$. Based on previous work on the cognitive plausibility of co-citation measures in the context of conceptual graphs, we adopt six state-of-the-art network similarity measures [8]. In particular, we consider P-Rank, a generic co-citation algorithm [19]. As discussed in Section 2.1, fordifferent parameters, P-Rank is equivalent to earlier algorithms, including Co-citation [17], Amsler [21], Coupling [20] and SimRank [18] and rvs-SimRank [19]. In this context, we adopt a formulation of P-Rank in linear algebra [8], discussing in detail the meaning and impact of its parameters (K, λ and C).

P-Rank is a recursive measure of similarity, based on the combination of two recursive assumptions: (1) two entities are similar if they are referenced by similar entities; and (2) two entities are similar if they reference similar entities. P-Rank is calculated iteratively, choosing a number of iterations $K \in [1, \infty)$. The higher K, the better the approximation of the theoretical solution to P-Rank. In this context, C is the P-Rank decay factor $\in (0, 1)$. Coefficient λ is the P-Rank in-out balance constant, in the interval $[0, 1]$. When $\lambda = 1$, only the incoming links are considered, and when $\lambda = 0$, only the outgoing links are included in the computation. Hence, we define s_{net} as follows:

$$s_{net}(a, b) = \lim_{k \to \infty} \mathbf{R}_k(a, b) \tag{1}$$
$$\mathbf{R}_k = C(\lambda \cdot \mathbf{T}_i \mathbf{R}_{k-1} \mathbf{T}'_i + (1 - \lambda) \cdot \mathbf{T}_o \mathbf{R}_{k-1} \mathbf{T}'_o) + \mathbf{\Theta}$$

where K is the P-Rank maximum iterations ($K \in [1, \infty)$). Matrix $\mathbf{R}_k$ is a P-Rank score matrix at iteration k. Matrix $\mathbf{T}_i$ is a transition matrix of $\mathbf{G}$ constructed on $I(a)$. In addition, $\mathbf{T}_o$ is the transition matrix of $\mathbf{G}$ constructed on $O(a)$, and $\mathbf{\Theta}$ is a diagonal matrix, so that $\forall k$, when $a = b$, $\mathbf{\Theta}(a, b) + \mathbf{R}_k(a, b) = 1$. All P-Rank iterations with $k > 0$ can be expressed as a series of iterations converging to the theoretical similarity score. Based on the optimization devised by Yu *et al.* [36], the computational complexity of this measure has the upper bound $O(n^3 + Kn^2)$.

3.2. Lexical Similarity (s_{lex})

The purpose of similarity function $s_{lex}(a, b)$ is the quantification of the semantic similarity of two text segments l_a and $l_b \in L$, which represent the lexical definition of nodes a and b in a GKG. Each concept is associated with a set of definitional terms $t_{a1}...t_{an}$ that describe the concept. If l_a or l_b are empty, $s_{lex}(a, b)$ is undetermined, and NLS has to rely only on s_{net}. In order to compute s_{lex} between two segments of text, we adopt the knowledge-based technique that we developed in our previous work [9]. The basic intuition behind this lexical similarity measure is that similar terms are described using similar terms. This bag-of-words measure computes the semantic similarity of two terms $s(a, b)$ based on input parameters $\{POS, C, sim_t, sim_v\}$: a part-of-speech (POS) filter, which consists of a set of POS tags (e.g., nouns and verbs); a corpus C; a term similarity function sim_t; and a vector similarity function sim_v. The four steps of the similarity algorithm are as follows:

1. Given two concepts a and b, lemmatize and POS-tag their terms in labels l_a and l_b.
2. Construct semantic vectors $\vec{a}$ and $\vec{b}$, based on definitional terms having POS contained in the POS filter. For each definitional term t, retrieve weights w_t from corpus C. A common approach to computing the weight of the definitional terms is the term frequency-inverse document frequency (TF-IDF). A relatively infrequent term in corpus C is expected to bear a higher weight than a frequent one.
3. Construct matrices M_{ab} and M_{ba}. Each cell of these similarity matrices contains a term similarity score $sim_t(t_{ai}, t_{bj})$. In principle, any term-to-term semantic similarity measure might be adopted as sim_t (see, for example, Table 1).
4. Compute similarity score $s_{lex}(a, b)$ from the similarity matrices using vector similarity sim_v, based on paraphrase-detection techniques, such as those by Corley and Mihalcea [34] or Fernando and Stevenson [37].

Having constructed the semantic vectors $\vec{a}$ and $\vec{b}$ and the matrices M_{ab} and M_{ba}, the vector-to-vector similarity sim_v in Step 4 deserves particular attention. First, an asymmetric similarity measure of semantic vectors $sim'_v(\vec{a}, \vec{b})$ can be formalized as follows:

$$sim'_v(\vec{a}, \vec{b}) = \sum_{i=1}^{|\vec{a}|} w_{ai} \cdot \hat{s}(t_{ai}, \vec{b}, M_{ab}), \quad sim'_v(\vec{b}, \vec{a}) = \sum_{i=1}^{|\vec{b}|} w_{bi} \cdot \hat{s}(t_{bi}, \vec{a}, M_{ba}) \tag{2}$$
$$sim'_v(\vec{a}, \vec{b}) \neq sim'_v(\vec{b}, \vec{a}), \quad sim'_v(\vec{a}, \vec{b}) \in [0, 1]$$

where function $\hat{s}$ returns a similarity score between a definitional term and a semantic vector, based on a similarity matrix. Two functions can be adopted as $\hat{s}$: either $\hat{s}_{com}$ (based on Corley and Mihalcea [34]) or $\hat{s}_{fes}$ (based on Fernando and Stevenson [37]). Finally, a symmetric measure $s_{lex} \in [0, 1]$ can be easily obtained from sim'_v as the average of $sim'_v(a, b)$ and $sim'_v(b, a)$. This knowledge-based approach relying on semantic vectors enables the computation of the lexical similarity in NLS. In terms of computational complexity, the upper bound of this measure is $O(n^3)$. As shown in the next section, to obtain a more plausible measure of similarity in GKGs, this component of semantic similarity can be combined with the network similarity.

3.3. Hybrid Similarity (s_{hyb})

In general, the limitations of computational approaches to the same problem can be overcome by combining them into an appropriate hybrid measure. In a GKG, some concepts might be situated in a densely-connected area of the network, while having sketchy labels. By contrast, other concepts can be poorly linked, but have richer lexical labels. This phenomenon sets upper bounds for network and lexical similarity, limiting the overall cognitive plausibility of the similarity measures.

Considering two concepts a and b in graph G, we have defined a network similarity measure $s_{net}(a, b)$ and a lexical similarity measure $s_{lex}(a, b)$. Both measures quantify the concept similarity with a real number in the interval $\Re \in [0, 1]$, where 0 means minimum similarity and 1 maximum similarity. In order to obtain a combined measure of similarity $s_{hyb}(a, b)$, we define two combination strategies: score combination (s_{sc}) and rank combination (s_{rk}). The score combination s_{sc} consists of the linear score combination of network and lexical similarities, weighted by a combination factor $\alpha \in [0, 1]$:

$$s_{sc}(a, b) = \frac{\alpha \cdot s_{net}(a, b) + (1 - \alpha) \cdot s_{lex}(a, b)}{2} \tag{3}$$

The rank combination s_{rk}, on the other hand, is the linear combination of the pair rankings, normalized on the cardinality of the pair set:

$$\begin{aligned} rk_{comb}(a, b) = \alpha \cdot rk(s_{net}(a, b)) + (1 - \alpha) \cdot rk(s_{lex}(a, b)) \\ s_{rk}(a, b) = \frac{|P| - rk_{comb}(a, b)}{|P| - 1} \\ rk_{comb} \in [1, |P|], \;\; s_{rk} \in [0, 1] \end{aligned} \tag{4}$$

where rk is a ranking function, P a set of concept pairs and α is the combination factor. While s_{sc} is a continuous function, s_{rk} is discrete. For example, in a set P of ten pairs, a pair of concepts (a, b) can have $s_{net} = 0.7$, resulting in $rk(s_{net}) = 3$ in the pair set. The lexical score $s_{lex} = 0.45$ might correspond to $rk(s_{lex}) = 8$. Fixing the value of α to 0.5, the score combination is $s_{sc} = 0.57$. The rank combination amounts to $rk_{comb} = 5.5$; therefore, $s_{rk} = 0.5$. The next section describes an empirical evaluation of NLS in a real-world scenario.

4. Evaluation

In this section, NLS is evaluated in a real-world scenario. The key purpose of this evaluation is the validation of the intuition underlying NLS: the complementary nature of network and lexical similarity in GKGs. These empirical results indicate that the hybrid measure can overcome the limitations of network and lexical measures. As ground truth, we selected the OSM Semantic Network, a GKG, and a corresponding dataset of human-generated similarity judgments, described in the next section. In the evaluation, we show that existing WordNet-based similarity measures are not sufficient to compute the semantic similarity in the context, and we analyze in detail the performance of the two components of NLS, assessing the superior plausibility of the hybrid measure. These results are then compared with those reported in our previous work, against the dataset defined in [11] to evaluate the matching-distance similarity measure (MDSM), a similarity measure for geographic concepts.

4.1. Ground Truth

As an evaluation testbed for NLS, we selected a GKG, the OSM Semantic Network [8]. This GKG contains a machine-readable representation of geographic concepts, extracted from the crowdsourced cartographic project OpenStreetMap. For example, the concept canal is represented by a vertex linked to concepts waterway and river (http://github.com/ucd-spatial/OsmSemanticNetwork). To date, the network contains about 5000 geographic concepts, linked by 19,000 edges. The OSM Semantic Network is a suitable choice, because it consists of a graph containing inter-connected concepts and whose concepts are associated with lexical descriptions. To evaluate NLS, we adopted the cognitive-plausibility approach, *i.e.*, the similarity judgments generated by the measure are compared against judgments obtained from human subjects.

Table 2. Human-generated similarity scores (H_{sc}) and rankings (H_{rk}) on 50 concept pairs, with 0 ties.

Concept A	Concept B	H_{sc}	H_{rk}	Concept A	Concept B	H_{sc}	H_{rk}
motel	hotel	0.9	1	picnic site	stream	0.37	26
theater	cinema	0.87	2	city	railway station	0.33	27
public transport station	railway platform	0.81	3	heritage item,area	valley	0.29	28
basketball court	volleyball facility	0.78	4	car store/shop	cycling facility	0.27	29
floodplain	wetland	0.77	5	office building	academic bookstore	0.27	30
stadium	athletics track	0.76	6	canoe spot	hunting shop	0.24	31
tram way	subway	0.76	7	school	toy shop	0.22	32
bay	body of water	0.76	8	post box	town	0.21	33
art shop	art gallery	0.75	9	supermarket	surveillance camera	0.2	34
historic battlefield	monument	0.67	10	arts center	currency exchange	0.16	35
restaurant	beverages shop	0.65	11	ambulance station	city	0.15	36
historic castle	city walls	0.64	12	shelter	agricultural field	0.15	37
administrative office	town hall	0.62	13	bed and breakfast	school building	0.14	38
tower	lighthouse	0.62	14	panoramic viewpoint	race track	0.12	39
police station	prison	0.61	15	football pitch	corporate office	0.11	40
canal	dock	0.59	16	beauty parlor	fire station	0.09	41
glacier	body of water	0.56	17	fashion shop	swimming spot	0.08	42
church	historic ruins	0.53	18	vending machine	gate	0.08	43
barracks	shooting range	0.51	19	city suburb	antiques furniture shop	0.07	44
mountain hut	hilltop, mountaintop	0.49	20	community center	stream	0.06	45
industrial land use	landfill	0.44	21	water ski facility	office furniture shop	0.05	46
swimming pool	water reservoir	0.42	22	interior decoration shop	tomb	0.05	47
managed forest	lone, significant tree	0.4	23	greengrocer	aqueduct	0.03	48
sea	island	0.39	24	political boundary	women's clothes shop	0.02	49
speed bump	car park	0.39	25	nursing home	continent	0.02	50

As a set of human psychological judgments, we selected the geo-relatedness and similarity dataset (GeReSiD) (http://github.com/ucd-spatial/Datasets) [10]. This dataset provides a set of human-generated similarity scores H_{sc} on 50 concept pairs rated by 203 human subjects, then ranked as H_{rk}, covering in total 97 concepts. Because semantic relatedness is outside the scope of this study, we considered only semantic similarity judgments. Following Resnik [16], we consider the upper bound for the cognitive plausibility of a computable measure to be the highest correlation obtained by a human rater with the dataset's means (Spearman's $\rho = 0.93$). In other words, this upper bound represents the empirical best results that human subjects obtained when rating the similarity of the concept pairs. Table 2 includes all

50 concept pairs, with the similarity score and ranking assigned by the human subjects, utilized in the next sections as ground truth.

4.2. WordNet-Based Experiment

This experiment aims at investigating the cognitive plausibility of WordNet-based similarity measures when applied directly to the concepts contained in GeReSiD. In order to evaluate the WordNet similarity measures directly on the concepts, the 97 OpenStreetMap concepts contained in GeReSiD were manually mapped to the corresponding WordNet synsets. The ten WordNet-based measures, summarized in Table 1, were computed on the 50 pairs. The resulting correlations of these similarity scores with the GeReSiD human scores obtain correlations with human similarity in the range $[0.53, 0.18]$. While some measures obtained relatively high plausibility (e.g., hso, $\rho = 0.53$), others resulted in weak correlations, showing very low cognitive plausibility. The statistically-significant results at $p < 0.05$ indicate ρ in the interval $[0.33, 0.53]$. The top performing measures are hso, vector and vectorp, obtaining $\rho \in [0.43, 0.53]$. The other measures obtain a considerably lower cognitive plausibility ($\rho < 0.34$), indicating no convergence towards the human-generated dataset. This experiment shows the inadequacy of WordNet-based measures applied directly to this GKG and the need for a more plausible measure.

4.3. Evaluation of Network Similarity

This section reports on the evaluation we conducted to assess the network component of NLS s_{net}. In order to evaluate the cognitive plausibility of co-citation measures applied to GKGs, an experiment was set up following and extending the approach that we adopted in [8]. The scores generated by the co-citation algorithms were compared with the similarity scores of the 50 pairs contained in GeReSiD, assessing their cognitive plausibility.

Network experiment setup. As discussed in Section 3.1, the recursive co-citation algorithm P-Rank includes a number of co-citation algorithms [19], including, among others, Coupling [20], Amsler [21] and SimRank [18]. To explore the performance of these network similarity measures, the following P-Rank parameters were selected:

- λ (P-Rank in-out link balance): 11 discrete equidistant levels $\in [0, 1]$.
- C (P-Rank decay constant): nine discrete equidistant levels $\in [0.1, 0.9]$. $C = 0.95$ was also included, being the optimal value for the domain [8].
- K (P-Rank iterations): 40 P-Rank iterations.

These parameters resulted in 4400 unique combinations of λ, C and K. The similarity scores were then obtained for the 50 concept pairs in GeReSiD, applying P-Rank for all of the 4400 combinations. The resulting 4400 sets of similarity scores were subsequently compared with the similarity scores of GeReSiD. The tie-corrected Spearman's rank correlation coefficient ρ was utilized to assess the correlation between machine and human-generated scores, on the 50 pair rankings without ties.

Network experiment results. The experiment resulted in 4400 correlations between co-citation similarity scores on the OSM Semantic Network and the corresponding similarity scores in GeReSiD, with $p < 0.01$ in all cases. All of the correlation tests were conducted on the 50 concept pairs, with a number of ties varying from zero to nine, 2.3 on average. In order to identify general trends in the results, the correlations are grouped by the three P-Rank parameters. As K increases, the similarity scores are closer to the theoretical, asymptotic value of P-Rank. In the results, the correlations quickly converge with $K \in [1, 10]$, followed by a slow decline in the interval $[11, 20]$, with $K > 20$, the correlations remain stable, around the mean $\rho = 0.62$, with standard deviation (SD) equal to 0.1.

The constant C determines how fast the similarity decays during the iterations. When $C \rightarrow 0$, the decay is fast, while $C \rightarrow 1$ implies a slow decay. For all of the values of C, the average correlation remains in the range $[0.55, 0.62]$, with $SD = 0.11$. Low values of C ($[0.1, 0.4]$) correspond to the lowest plausibility in the experiment ($\rho < 0.65$). The best results are obtained when $C \in [0.5, 0.9]$, with a peak at $C = 0.8$ ($\rho = 0.62$) and a drop when $C = 0.95$. The third parameter that influences the results of P-Rank is λ, the balance between in- and out-links in the semantic network. When $\lambda = 0$, only the out-links are considered, while $\lambda = 1$ includes only in-links.

Figure 2 shows the impact of λ on the cognitive plausibility of P-Rank. Each point on the plot represents the average of 410 correlations, falling in the range $[0.48, 0.65]$, with $SD \approx 0.1$. The performance of the algorithms improves steadily as λ moves from zero to 0.9, with a peak at $\lambda = 0.9$ (mean $\rho = 0.69$). When $\lambda = 1$, the performance decreases suddenly ($\rho = 0.63$), indicating that out-links provide useful information. Hence, focusing on the best approximations to the theoretical value of P-Rank ($K = 40$), the most plausible results against GeReSiD are located in the intervals $C \in [0.5, 0.8], \lambda \in [0.8, 0.9]$. In this region, the mean correlation with the human rankings reaches $\rho = 0.73$. Table 3 summarizes the results of this evaluation, comparing the cognitive plausibility of the s_{net} algorithms against GeReSiD, including the results with the MDSM evaluation dataset from [8].

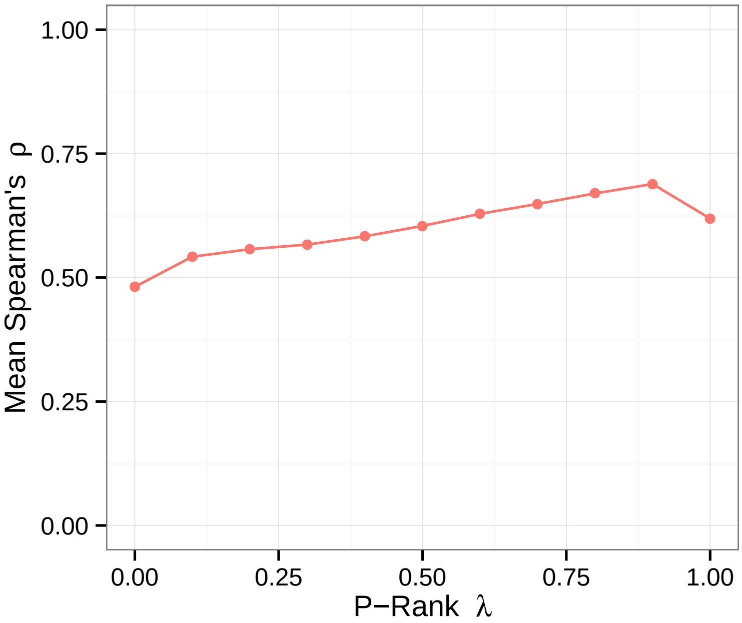

Figure 2. Experiment results grouped by P-Rank in-out link factor λ.

Network dataset comparison. Although the GeReSiD results show substantial agreement with the MDSM evaluation dataset, differences between the two datasets exist. The optimal performance of P-Rank in GeReSiD is obtained with parameters $C = 0.8, \lambda = 0.9$. By contrast, the MDSM evaluation dataset is best approximated when $C = 0.9, \lambda = 1$, corresponding to the SimRank algorithm. The plausibility of P-Rank suddenly drops when $\lambda = 1$ in GeReSiD, which does not occur in the MDSM evaluation dataset. This difference is due to the limited information problem that affects SimRank, as Zhao *et al.* [19] pointed out. As SimRank only relies on in-links, vertices that have only out-links cannot obtain a similarity score. The different coverage in the two datasets can also help explain these differences. While the MDSM evaluation dataset contains 29 concepts, GeReSiD covers 97 OpenStreetMap concepts, including more concepts affected by the limited information problem.

Table 3. Cognitive plausibility of network similarity measures s_{net}. *Sim* stands for similarity. MDSM results from [8]. * Best performance.

K	λ	C	Network	MDSM	GeReSiD
$[1, \infty)$	$[0, 1]$	$(0, 1)$	Measure	Sim ρ	Sim ρ
1	0	–	Coupling [20]	0.55	0.5
1	0.5	–	Amsler [21]	0.67	0.53
1	1	–	Co-citation [17]	0.72	0.61
10	0	0.9	rvs-SimRank [19]	0.57	0.46
10	0	0.5	–	0.57	0.5
10	0	0.1	–	0.6	0.51
10	0.5	0.9	P-Rank [19]	0.76	0.64
10	0.5	0.5	–	0.73	0.62
10	0.5	0.1	–	0.67	0.6
40	0.9	[0.5,0.8]	–	–	[0.72,0.73 *]
10	1	0.9	SimRank [18]	0.85 *	0.65
10	1	0.5	–	0.78	0.64
10	1	0.1	–	0.75	0.64

Network similarity limitations. Although optimal parameters lead to strong correlation for similarity ($\rho \approx 0.7$), it is beneficial to assess the cases in which the network similarity measures show a considerable discrepancy with the human-generated rankings. When $K = 40$, $C = 0.8$ and $\lambda = 0.9$, concept pair <*arts center, bureau de change*> is ranked 35th in the set of 50 pairs by the human subjects, while the pair is ranked sixth by P-Rank. This wide gap is due to the high structural similarity of the two concepts, which are both linked to the key *amenity*, and are not densely linked to other concepts that might help the algorithm reduce their semantic similarity. The opposite case arises with two pairs <*city, railway station*> and <*heritage item, valley*>, which are ranked respectively 27th and 28th by the human subjects and are ranked 44th and 45th by P-Rank. These weak relations are not captured by the link structure in the OSM Semantic Network, and therefore, P-Rank fails to find any similarity between the pairs.

4.4. Evaluation of Lexical Similarity

This section discusses the evaluation that we have conducted on the lexical similarity component s_{lex} of NLS, outlined in Section 3.2, using GeReSiD as ground truth. This approach consists of extracting vectorial representations of the lexical definitions and then comparing them using term-to-term semantic similarity measures. The overall label-to-label similarity measure is subsequently obtained by combining the term similarity matrix using paraphrase detection techniques.

Lexical experiment setup. The experiment consists of a set of 180 combinations of the technique's four input parameters $\{POS, C, sim_t, sim_v\}$, detailed in Table 4. All of the rankings generated in this phase contained no ties and were compared with the GeReSiD using Spearman's ρ.

Table 4. Lexical experiment setup: resources included as input parameters. POS, part-of-speech.

Param	#	Description
POS	3	POS filters (*NN, VB, NN VB*). Adjectives (*JJ*) were initially included, but most measures sim_t are designed to handle only nouns and verbs, making a direct comparison difficult.
C	3	We collected two corpora: the OSM Wiki website and a set of random news stories from the newspaper, *Irish Independent*. Both contain about 2.5 M words. The *Null*corpus corresponds to constant weights, *i.e.*, a constant $w > 0$.
sim_t	10	The term-to-term similarity function sim_t is used to construct the similarity matrices needed to compute the similarity: *hso*, *jcn*, *lch*, *lesk*, *lin*, *path*, *res*, *vector*, *vectorp* and *wup*(see Table 1).
sim_v	2	Two vector-to-vector similarity measures, originally developed to detect paraphrases, were included: *com* [34] and *fes*[37]
Total	180	$\lvert POS\rvert \cdot \lvert C\rvert \cdot \lvert sim_t\rvert \cdot \lvert sim_v\rvert$

Lexical experiment results. The results are summarized in Table 5, which for each parameter reports median, quartiles and maximum ρ. As the distributions of ρ for the algorithm parameters tend to be heavily skewed, we adopt the median $\tilde{\rho}$ as a robust estimator of central tendency, reporting the 25% and 75% quartiles for each parameter. As already noted in relation to the results in [9], verbs used in isolation (POS = *VB*) do not show correlation with the human dataset, resulting in $\rho \in [0.01, 0.16]$, with $p > 0.1$. Similar issues apply to the *fes*vector-to-vector measure, which obtained $\tilde{\rho} = 0.26$, with $p > 0.05$. Hence, these non-significant results were excluded from the analysis. For all of the other cases, the correlations were statistically significant with $p < 0.001$.

Overall, the lexical component of the NLS approach to computing semantic similarity in a GKG obtains a median $\tilde{\rho} = 0.61$, with the upper bound being $\rho = 0.74$. The four parameters that influence the algorithm results are $\{POS, C, sim_t, sim_v\}$. The vector-to-vector measure sim_v determines the strategy to compute the similarity of semantic vectors. While *fes* did not show satisfactory cognitive plausibility, *com* obtained more plausible results. The POS filter selects the terms to be included in the semantic vectors. Excluding the analysis of verbs in isolation (*VB*), *NN* and *NN VB*show a very close cognitive plausibility ($\tilde{\rho} = 0.61$). The text corpus C is utilized to assign semantic weights to the terms. The cognitive plausibility obtained by the *Null* and OSM Wiki corpora is largely comparable

($\tilde{\rho} = 0.58$). By contrast, the corpus extracted from the Irish Independent, containing news stories, outperforms the other corpora, resulting in a higher cognitive plausibility ($\tilde{\rho} = 0.64$), showing that the non-domain-specific corpus supports the computation better than a domain-specific corpus.

Table 5. Results of the lexical similarity experiment. MDSM results from [9]. * Best performance. GeReSiD, geo-relatedness and similarity dataset.

		MDSM	**GeReSiD Sim**		
Param Name	**Param Value**	**Median** $\tilde{\rho}$	**Median** $\tilde{\rho}$	**25%–75% Quartiles**	**Max** ρ
sim_v	com	0.7	0.61	0.56 0.64	0.74
	fes	0.66	–	– –	–
POS	NN	0.68	0.61 *	0.56 0.64	0.74 *
	NN VB	0.7	0.61 *	0.56 0.64	0.73
	VB	–	–	– –	–
C	Irish Indep	0.7	0.64 *	0.62 0.72	0.74
	Null	0.7	0.58	0.54 0.62	0.64
	OSM Wiki	0.67	0.58	0.52 0.62	0.65
sim_t	vector	0.62	0.64 *	0.64 0.71	0.74 *
	path	0.7	0.64 *	0.64 0.71	0.73
	lch	0.74	0.62	0.62 0.7	0.73
	hso	0.71	0.61	0.61 0.66	0.71
	wup	0.74	0.6	0.57 0.62	0.66
	res	0.72	0.6	0.59 0.62	0.64
	lesk	0.48	0.56	0.56 0.62	0.64
	vectorp	0.57	0.54	0.52 0.6	0.64
	jcn	0.75	0.5	0.49 0.55	0.59
	lin	0.67	0.48	0.45 0.5	0.56
all	–	0.69	0.61	0.56 0.64	0.74

The fourth parameter, which has a high impact on the results, is the term-to-term measure sim_t. Measures *vector*, *path*, *lch* and *hso* fall in the top tier, with an upper bound $\rho \geq 0.7$ and a median $\tilde{\rho} > 0.6$. All of the other measures perform in a less satisfactory way, with a lower median in the interval $[0.48, 0.6]$ and an upper bound $\rho \in [0.56, 0.66]$. After the top cluster of these four term-to-term measures, the performance drops visibly, reaching a minimum with *lin* (median ≈ 0.47, upper bound ≈ 0.55). The other measures (*wup*, *res*, *lesk*, *vectorp* and *jcn*) fall between the top four, reaching intermediate results. The lexical similarity measures s_{lex} outperform the basic WordNet-based measures, with an upper bound $\rho = 0.74$). The top performance is reached with the following parameters: POS = *NN*, C = *Irish Indep*, sim_v = *com*, sim_t = $\{$*vector*, *path*, *lch*, *hso*$\}$). In such cases, the cognitive plausibility ρ falls in the interval $[0.61, 0.74]$, showing a statistically-significant strong correlation with GeReSiD.

Lexical dataset comparison. Table 5 includes the median $\hat{\rho}$ that we obtained with the MDSM evaluation dataset in [9]. The cognitive plausibility obtained in these two evaluations shows common trends, but also a divergence for certain parameters. This fact is consistent with the evaluation of network similarity, in which co-citation approaches performed better on the MDSM evaluation dataset than on GeReSiD. This difference is mostly due to the structure and coverage of the MDSM evaluation dataset (29 concepts structured in five sets) and GeReSiD (97 concepts in one set). While the overall trends in the two experiments on lexical similarity are consistent, the effect of individual parameters $\{POS, C, sim_t, sim_v\}$ varies.

In particular, sim_t have a major impact on the cognitive plausibility of the algorithm. A high variability can be noticed between the two experiments, which is not uncommon in the literature on semantic similarity. In a study by Budanitsky and Hirst [38], measures *jcn*, *hso*, *lin*, *lch* and *lesk* obtain very different cognitive plausibility against two well-known similarity datasets. The measures that reach the top overall performance are *lch*, *path*, *vector* and *hso*, with upper bounds in the range $[0.72, 0.75]$. The other measures rank lower, falling in the interval $[0.62, 0.69]$. It is possible to notice that, although more complex measures can obtain optimal results in certain contexts, simpler shortest path-based measures, such as *path* and *lch*, tend to perform more reliably across the two datasets.

Lexical similarity limitations. Although s_{lex} can reach high plausibility, specific cases show high discrepancy with the human-generated similarity judgments in the set of 50 concept pairs in GeReSiD. Focusing on the best case (POS = *NN*, C = *Irish Indep*, sim_v = *com*, sim_t = *vector*, with $\rho = 0.74$), it is possible to observe that the pair <*sea, island*> is ranked 24th by human subjects and eighth by the algorithm. The definitions of these two concepts have large lexical overlap, but they are highly related (eighth in the relatedness ranking) and not similar. In this case, the algorithm mistakes relatedness for similarity.

Furthermore, <*battlefield, monument*> is ranked 10th by the human subjects, and only 36th by the algorithm. The concepts' labels share only one term (*military*), and the other terms do not increase their similarity. Analogously, the similarity of <*industrial land use, landfill*> is underestimated, as it is ranked 21st by humans and 47th by the algorithm. The reason for this wide mismatch lies in the fact that the label of landfill is extremely short ("where waste is collected, sorted or covered over") and does not contain terms that would allow the algorithm to capture some degree of similarity with the context of industrial production and waste processing. These limitations can be overcome by combining s_{net} and s_{lex} into a hybrid measure, as shown in the next section.

4.5. Evaluation of Hybrid Similarity

As stated in Section 3.3, two methods can be used to combine s_{net} and s_{lex} into a hybrid measure: a score combination s_{sc} and a rank combination s_{rk}. This section describes an empirical evaluation of these two combination techniques, showing that the cognitive plausibility of such hybrid measures is generally higher than the individual network and lexical measures, supporting the intuition behind NLS.

Hybrid experiment setup. To explore the effectiveness of score and rank combination methods, a cognitive plausibility experiment was set up using GeReSiD. The most plausible cases were selected for network s_{net} and lexical measures s_{lex}, based on the empirical results shown in Sections 4.3 and 4.4.

As we are interested in assessing whether the combination methods are able to improve the results at the top of the range, the selection is restricted to the top 30 cases for both approaches, as a representative sample of the network and lexical measures. These top cases are not statistical outliers, but accurately reflect general trends in the empirical evidence collected in the aforementioned experiments. The experiment was set up with the following input parameters:

- Combination methods: score combination s_{sc} and rank combination s_{rk}.
- Combination factor α: ten discrete equidistant levels $\in [0, 1]$. When $\alpha = 0$, only the lexical measure is considered; $\alpha = 1$, on the other hand, corresponds to the network measure.
- Network similarity s_{net}: 30 most cognitively plausible cases when compared with GeReSiD.
- Lexical similarity s_{lex}: 30 most cognitively plausible cases when compared with GeReSiD.

For each value of α, each case of s_{net} and s_{lex} were combined through s_{sc} and s_{rk}. This resulted in the cognitive plausibility of 18,000 hybrid measures on the 50 concept pairs of the GeReSiD, with $p < 0.001$ for all of Spearman's correlation tests, with no ties in the rankings. A hybrid measure is considered successful if it outperforms both its components s_{net} and s_{lex}, *i.e.*, the cognitive plausibility of the hybrid measure is strictly greater than network and lexical similarity, formally $\rho_{hyb} > \rho_{net} \wedge \rho_{hyb} > \rho_{lex}$. If the hybrid measure is lower or equal to any of its components, it has failed.

Table 6. Cognitive plausibility of NLS. Max ρ is the upper bound obtained by an approach. *net*: network measure; *lex*: lexical measure; *hyb*: hybrid measure. * Best performance. For all Spearman's tests, $p < 0.001$.

		Score Comb s_{sc}		**Rank Comb s_{rk}**	
	α	ρ	**Success %**	ρ	**Success %**
lex	0	0.74	–	0.74	–
hyb	0.2	0.79	74.4	0.79	73.1
hyb	0.4	0.81	87.5	0.83	100.0
hyb	0.5	0.82 *	91.9	0.84 *	100.0
hyb	0.6	0.81	95.6	0.83	100.0
hyb	0.8	0.8	96.9	0.79	86.9
net	1	0.73	–	0.73	–

Hybrid experiment results. Clear patterns emerge from the experiment results. Hybrid measures, combining network and lexical similarity, show a consistent advantage over their network and lexical components. The ranking combination s_{rk} performs consistently better than the score combination s_{sc}, obtaining higher plausibility and success rate. Table 6 summarizes the experiment results, contrasting the upper bound of ρ obtained by *net* and *lex* measures in isolation, and *hyb* when combined. The cognitive plausibility of hybrid measures is substantially greater than the individual measures, with a peak at $\rho = 0.84$ when $\alpha = 0.5$. This empirical evidence points out that the optimal value of α tends to fall in the interval $[0.4, 0.6]$, drawing information evenly from the network and lexical components.

The success rate, expressed as a percentage, indicates in how many cases a hybrid measure outperformed both of the individual measures. As it is possible to notice in Table 6, when $\alpha \in [0.4, 0.6]$, the success rate is very high, in the interval $[87.5\%, 100\%]$. In particular, the rank combination s_{rk} outperforms all individual measures (100%). High success rates are also observable when $\alpha \in (0, 0.4)$, with an average success rate of 82.9%. At the other end of the spectrum ($\alpha \in (0.6, 1)$), the average success rate is 75%. In none of the cases under consideration was a hybrid measure lower than both its components.

The success rates reported in Table 6 show that, overall, both components strongly contribute to the cognitive plausibility of NLS. In particular, when using ranking combination s_{rk} with optimal values of α, the hybrid measures obtain a success of >89%. The performance of NLS is depicted in Figure 3, highlighting the impact of α on the cognitive plausibility, adopting the two combination techniques (s_{sc} and s_{rk}). The roughly symmetrical bell-shaped curves in the figure display the benefit of the hybrid measures ($\alpha \in (0, 1)$) over the individual measures, at the extremes of the horizontal axis ($\alpha = 0$ corresponds to lexical measures, $\alpha = 1$ to network measures).

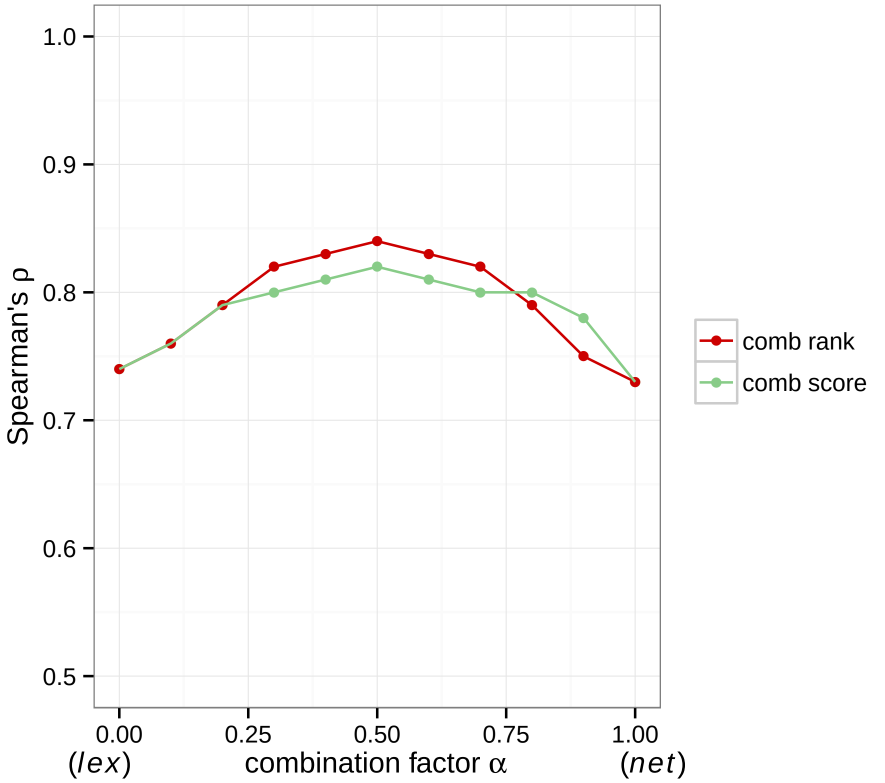

Figure 3. Cognitive plausibility of hybrid measures. *comb rank*: rank combination s_{rk}; *comb score*: score combination s_{sc}; $\alpha \in [0, 1]$.

Hybrid similarity limitations. Considering the best hybrid measures ($\alpha = 0.5$), it is possible to observe changes with respect to the rankings generated by individual measures. Only in one case do the hybrid measures fail to improve on the previous measures, ranking <*sea, island*> fifth (see Table 2 for a comparison with human rankings). In all of the other cases discussed above, the hybrid measures provide more cognitively plausible rankings: <*arts center, bureau de change*> (15th), <*city, railway station*> (19th), <*heritage item, valley*> (28th), <*battlefield, monument*> (29th), and <*industrial land use, landfill*> (44th). In summary, the hybrid measures cannot fully overcome the limitations intrinsic to the data source, but they succeed, on average, in bringing the rankings closer to human judgments. Based on this body of empirical evidence, the hybrid approach is the most suitable to compute semantic similarity in GKGs.

5. Conclusions

In this article, we described network-lexical similarity measure (NLS), a measure designed to capture the similarity of concepts in GKGs, knowledge-representation structures used to represent concepts and their relations. The evaluation on the OSM Semantic Network confirmed the benefits of combining network similarity and lexical similarity into a hybrid measure, obtaining higher cognitive plausibility. Compared with the upper bounds for network measures ($\rho = 0.73$) and lexical measures ($\rho = 0.74$), hybrid measures reach a considerably higher upper bound ($\rho = 0.84$). In order to provide practical guidelines, Table 7 summarizes the optimal results of the network, lexical and hybrid measures.

Although NLS obtains high cognitive plausibility, overcoming the intrinsic issues of network and lexical similarities, some limitations remain to be addressed in future research. The network measures s_{net} that we included in this study have cubic complexity, and substantial spatio-temporal optimization is needed to apply them to large GKGs [39]. In relation to s_{lex}, the paraphrase-detection techniques utilized in the lexical component need optimizations to be applicable on a very large scale. In addition, WordNet has limitations in coverage and biases. The method described in s_{lex} utilizes a bag-of-words model for the terms in the lexical descriptions. However, in many cases, the most important terms tend to be located at the beginning of the descriptions, and taking the term order into account might improve the results, especially in cases of very long and noisy lexical definitions. Furthermore, fully-corpus-based measures could be utilized in s_{lex} to overcome NLS to increase its coverage, at the expense of some precision.

From a more cognitive viewpoint, the main limitation of NLS lies in the lack of a precise context for the computation of the similarity measure, as illustrated by Keßler [40]. Other limitations affect the evaluation of cognitive plausibility that we have adopted in Section 4. Human subjects grasp semantic similarity intuitively, but the translation of a similarity judgment into a discrete number can be highly subjective, limiting the inter-rater agreement and the generalizability of the results [41]. In this article, we evaluated NLS on its ability to simulate human judgments on the entire range of semantic similarity, *i.e.*, from very similar to very dissimilar concepts. However, many similarity applications need specifically the top-k most similar concepts to a given concept, rather than the least-similar concepts. Given that no cognitive plausibility evaluation is fully generalizable, robust evidence can only be constructed by cross-checking different evaluations. For example, complementary indirect evaluations could focus on specific similarity-based tasks, such as word sense disambiguation and information retrieval. The approach to semantic similarity adopted in NLS can be extended to computing measures of relatedness, which have vast applicability [38].

Our evaluation focused on the OSM Semantic Network as a GKG. While this semantic network shows typical characteristics of GKGs [8], it is restricted to a very specific domain. Others GKGs suitable for the evaluation of NLS might be YAGO, DBpedia and other linked open datasets [5]. Moreover, the general text corpus we used presents a regional bias, and larger, more global corpora might further improve the results. However, cognitive plausibility evaluations on large, domain-independent GKGs are difficult to design, and a trade-off between domain-specificity and result reliability has to be considered. This future work will strengthen the role of NLS as a generic approach to tackling the challenge of the computation of semantic similarity, in ubiquitous GKGs, which increasingly contain valuable knowledge that complements that of traditional geographic datasets.

Table 7. Summary of optimal parameter values for network, lexical and hybrid measures.

	Network similarity (s_{net})
Balance λ	Incoming links indicate similarity more than outgoing links. Optimal $\lambda \in [0.9, 1]$.
Max K	Recursive co-citation algorithms reach high cognitive plausibility. SimRank [18] obtains $\rho = 0.65$, while P-Rank [19] reaches 0.73. Optimal $K > 20$.
Decay C	Slow decay is better than fast decay. Optimal $C \in [0.8, 0.9]$.
	Lexical similarity (s_{lex})
Corpus C	The text corpus extracted from the *Irish Independent* newspaper outperforms the domain-specific corpus.
POS filter	The nouns convey most semantic similarity. Verbs in isolation obtain low cognitive plausibility. When combined with nouns, verbs affect the results only slightly.
Term sim_t	Optimal term-to-term similarity measures: *lch* [25], *path* [23], and *vector* [29]. The other measures sim_t obtain lower plausibility.
Vector sim_v	Optimal vector similarity measure: *com* [34]. Measure *fes* [37] did not obtain statistically significant correlation.
	Hybrid similarity (s_{hyb})
Combination factor α	Both components, network and lexical, are needed to obtain cognitively plausible results. Optimal value of combination factor: $\alpha \in [0.4, 0.6]$
Combination method s	Optimal combination method: ranking combination s_{rk} outperforms direct score combination s_{sc}.

Acknowledgments

The research presented in this article was funded by a Strategic Research Cluster Grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

Author Contributions

Andrea Ballatore is the leading author of this work. He conceived of the core ideas and carried out the implementation and coding of the experiments as part of his PhD at University College Dublin, funded by Science Foundation Ireland. Michela Bertolotto and David C. Wilson were respectively supervisor and advisor of his PhD. They gave substantial contributions to the design and analysis of this work and to the drafting and critical review of the article.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Chein, M.; Mugnier, M. *Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs*; Springer: Berlin, Germary, 2008.
2. Heath, T.; Bizer, C. Linked data: Evolving the web into a global data space. *Synth. Lect. Semant. Web Theory Technol.* **2011**, *1*, 1–136.
3. Fellbaum, C. WordNet. In *Theory and Applications of Ontology: Computer Applications*; Poli, R., Healy, M., Kameas, A., Eds.; Springer: Berlin, Germany, 2010; pp. 231–243.
4. Goodchild, M. Citizens as sensors: The World of Volunteered Geography. *GeoJournal* **2007**, *69*, 211–221.
5. Ballatore, A.; Wilson, D.; Bertolotto, M. A survey of volunteered open geo-knowledge bases in the semantic web. In *Quality Issues in the Management of Web Information*; Pasi, G., Bordogna, G., Jain, L., Eds.; Intelligent Systems Reference Library: Berlin, Germany, 2013; Volume 50, pp. 93–120.
6. Purves, R.; Jones, C. Geographic information retrieval. *SIGSPATIAL Spec.* **2011**, *3*, 2–4.
7. Euzenat, J.; Meilicke, C.; Stuckenschmidt, H.; Shvaiko, P.; Trojahn, C. Ontology Alignment Evaluation Initiative: Six years of experience. *J. Data Semant. XV* **2011**, *6720*, 158–192.
8. Ballatore, A.; Bertolotto, M.; Wilson, D. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowl. Inf. Syst.* **2013**, *37*, 61–81.
9. Ballatore, A.; Bertolotto, M.; Wilson, D. Computing the semantic similarity of geographic terms using volunteered lexical definitions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2099–2118.
10. Ballatore, A.; Bertolotto, M.; Wilson, D. An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica* **2014**, *18*, 747–767.
11. Rodríguez, M.; Egenhofer, M. Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 229–256.
12. Budanitsky, A.; Hirst, G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the 2nd Meeeting of the North American Chapter of the Association for Computational Linguistics, Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA, USA, 2–7 June, 2001, pp. 29–34.
13. Turney, P. Similarity of semantic relations. *Comput. Linguist.* **2006**, *32*, 379–416.

14. Janowicz, K.; Keßler, C.; Schwarz, M.; Wilkes, M.; Panov, I.; Espeter, M.; Bäumer, B. Algorithm, implementation and application of the SIM-DL similarity server. In Proceedings of the GeoSpatial Semantics: Second International Conference, GeoS 2007, Mexico City, Mexico, 29–30 November 2007; Volume 4853, pp. 128–145.
15. Schwering, A. Approaches to semantic similarity measurement for geo-spatial data: A survey. *Trans. GIS* **2008**, *12*, 5–29.
16. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95. Morgan Kaufmann, Montreal, QC, Canada, 20–25 August 1995; Volume 1, pp. 448–453.
17. Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.* **1973**, *24*, 265–269.
18. Jeh, G.; Widom, J. SimRank: A measure of structural-context similarity. In Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–25 July 2002; pp. 538–543.
19. Zhao, P.; Han, J.; Sun, Y. P-rank: A comprehensive structural similarity measure over information networks. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, Hong Kong, China, 2–6 November 2009; pp. 553–562.
20. Kessler, M. Bibliographic coupling between scientific papers. *Am. Doc.* **1963**, *14*, 10–25.
21. Amsler, R. *Applications of Citation-Based Automatic Classification*; Technical Report 14; Linguistics Research Center: Austin, TX, USA, 1972.
22. Oliva, J.; Serrano, J.I.; del Castillo, M.D.; Iglesias, Á. SyMSS: A syntax-based measure for short-text semantic similarity. *Data Knowl. Eng.* **2011**, *70*, 390–405.
23. Rada, R.; Mili, H.; Bicknell, E.; Blettner, M. Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 17–30.
24. Wu, Z.; Palmer, M. Verbs semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL-94, Las Cruces, NM, USA, 27–30 June 1994; pp. 133–138.
25. Leacock, C.; Chodorow, M. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*; Fellbaum, C., Ed.; MIT Press: Cambridge, MA, USA, 1998; pp. 265–283.
26. Hirst, G.; St-Onge, D. Lexical chains as representations of context for the detection and correction of malapropisms. In *WordNet: An Electronic Lexical Database*; Fellbaum, C., Ed.; MIT Press: Cambridge, MA, USA, 1998; pp. 305–332.
27. Banerjee, S.; Pedersen, T. An adapted Lesk algorithm for word sense disambiguation using WordNet. In Proceedings of Third International Conference, CICLing 2002, Mexico City, Mexico, 17–23 February 2002; Volume 2276, pp. 117–171.
28. Jiang, J.; Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, ROCLING X, Taipei, Taiwan, 22–24 August 1997; Volume 1, pp. 19–33.

29. Patwardhan, S.; Pedersen, T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In Proceedings of the EACL 2006 Workshop Making Sense of Sense–Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy, 4 April 2006; Volume 1501, pp. 1–8.
30. Lin, D. An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, USA, 24–27 July 1998; Volume 1, pp. 296–304.
31. Ballatore, A.; Bertolotto, M.; Wilson, D. The semantic similarity ensemble. *J. Spat. Inf. Sci.* **2014**, doi: 10.5311/JOSIS.2013.7.128.
32. Landauer, T.; McNamara, D.; Dennis, S.; Kintsch, W. *Handbook of Latent Semantic Analysis*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2007.
33. Turney, P. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the 12th European Conference on Machine Learning, ECML'01, Freiburg, Germany, 5–7 September, 2001; Volume 2167, pp. 491–502.
34. Corley, C.; Mihalcea, R. Measuring the semantic similarity of texts. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, MI, USA, 30 June 2005; pp. 13–18.
35. Mihalcea, R.; Corley, C.; Strapparava, C. Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the Twenty-First National Conference on Artificial Intelligence, Boston, MA, USA, 16–20 July 2006; Volume 21, pp. 775–780.
36. Yu, W.; Lin, X.; Le, J. Taming computational complexity: Efficient and parallel simRank optimizations on undirected graphs. In Proceedings of the 11th International Conference on Web-Age Information Management, WAIM 2010, Jiuzhaigou Valley, China, 15–17 July 2010; Volume 6184, pp. 280–296.
37. Fernando, S.; Stevenson, M. A semantic similarity approach to paraphrase detection. In Proceedings of Computational Linguistics UK (CLUK 2008), 11th Annual Research Colloquium, Computational Linguistics UK, Oxford, UK, 18–20 March 2008; pp. 1–7.
38. Budanitsky, A.; Hirst, G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Comput. Linguist.* **2006**, *32*, 13–47.
39. Li, P.; Liu, H.; Yu, J.; He, J.; Du, X. Fast single-pair SimRank computation. In Proceedings of the SIAM International Conference on Data Mining, SDM2010, Columbus, OH, USA, 29 April–1 May 2010; pp. 571–582.
40. Keßler, C. Similarity measurement in context. In Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context, Roskilde, Denmark, 20–24 August, 2007; Volume 4635, pp. 277–290.
41. Ferrara, F.; Tasso, C. Evaluating the results of methods for computing semantic relatedness. In *Computational Linguistics and Intelligent Text Processing*; Gelbukh, A., Ed.; Springer: Berlin, Germany, 2013; Volume 7816, pp. 447–458.